Journal of Engineering Science and Technology Review 16 (1) (2023) 61 - 67

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Facial Expression Recognition Based on Improved Convolutional Neural Network

Liu Siyuan[1], Wang Libiao[2,*] and Zheng Yuzhen[1]

[1]*School of Mechanical and Energy Engineering, Zhejiang University of Science & Technology, Hangzhou 310023, China*
[2]*School of Intelligent Manufacture, Taizhou University, Taizhou 318000, China*

Received 9 September 2022; Accepted 22 February 2023

Abstract

The accuracy of traditional convolutional neural network (CNN) for facial expression recognition (FER) is not high. To improve detection accuracy, a face micro-expression recognition algorithm combining image deflection angle weighting, which achieves static and dynamic facial expression recognition, was proposed. First, face recognition was performed on the face to be measured based on the Haar features in the OpenCv library. Second, pre-processing, such as face position detection, face cropping, normalization, and data enhancement, was performed on the measured image to avoid irrelevant information interfering with the judgment. Third, convolutional neural network was used for FER, and the result of the linear weighting of expression labels measured by deflecting the face to be tested by multiple angles was used as final recognition result to improve the accuracy. Lastly, a camera was used for real-time judgments and static recognition on the CK+ data set. Results show that classifying difficult images in multiple combinations and building integrated models improve prediction accuracy. The recognition rate on the CK+ data set is 97.85%, which is an improvement of about 3% compared with the cross-connect LeNet-5 network algorithm, thereby verifying its feasibility and effectiveness. This study provides a good reference for the improvement of facial expression detection performance.

Keywords: Facial expression recognition (FER), CNN, Face detection, Deflection angle

1. Introduction

The American psychologist Ekman studied facial expressions and found six types of human facial expressions, each with different emotions: anger, surprise, disgust, enjoyment, fear, and sadness [1]. Facial expressions are an important way for humans to communicate non-verbal emotions. From changes in expressions, one can perceive people's emotions, feelings, temperament, and other psychological states. Therefore, facial expression recognition (FER) has become a major popular topic in the computer vision field. With the development of artificial intelligence (AI), the accuracy of FER is also improving rapidly, and FER systems are widely used in security monitoring, assisted driving, intelligent human–computer interaction (HCI), and behavioral science and medical research [2].

However, with the development of technology, facial expressions of the human face are the most direct and effective means to express emotions in the AI field. In the future, research on FER technology will be the main development direction of human–computer emotional interaction. Accordingly, providing immense challenges to the FER research is how to make computers better perceive face information under the influence of various factors (e.g., lighting, viewing angle, occlusion, age), reduce the computational overhead of pictures' additional information, and improve the recognition rate.

On the basis of the preceding issue, scholars have conducted extensive research on the network generating extensive redundant information in the FER process and improving the expression recognition rate [3–5]. However, there are still bottlenecks in the recognition rate problem and the coupling relationship between data feature extraction and its correlation. Therefore, an urgent problem that should be solved is how to accurately detect facial expressions and collect a large amount of sufficient training data to clarify the influencing factors that are nonlinearly coupled with facial expression data.

This study proposes a convolutional neural network (CNN) facial expression dynamic recognition model based on the angle transformation of images to be measured and manual feature extraction and classification optimization. The model is capable of detecting static pictures to be tested, as well as dynamic pictures and detecting FER in real time. The proposed system effectively prevents the use of high-definition photos and high-definition display devices (e.g., cell phones and PADs) to display photos and videos to deceive detection devices.

2. State of the art

At present, human expression has been studied extensively by scholars from different countries. Face expressions contain extensive personal behavior information and also have a wide range of applications in health care, Human–Computer Interaction (HCI) and security monitoring. Chen [6] proposed Weber local descriptor (WLD) features based on Weber's law to describe and extract features from the grayscale changes of pixel points in the neighborhood, with good robustness to illumination and noise. However, the description of image detail information is insufficient and does not substantially improve accuracy. Tan [7] used a local gradient feature calculation method with hierarchical

*E-mail address: wanglibiao@tzc.edu.cn
ISSN: 1791-2377
doi:10.25103/jestr.161.08

structure to achieve classification by support vector machines. However, some of the results were achieved in traditional expression recognition. Meanwhile, feature extraction is required human design and considerably relies on the understanding and definition of expression features. Alpaslan [8–9] proposed a hybrid local binary model based on the Hessian matrix and gravitational centrosymmetric LBP, which was fused using a cross-scale joint coding strategy, achieving good results in classification. However, manually defined features have difficulty in significantly describing the deformation of information from facial expression changes. To solve the problems caused by hand-designed features, Zhou [10] proposed an expression feature extraction method based on Gabor wavelet features and eye, nose, and mouth (ENM) differential weights. This method extracts features from different regions and adaptively weights them, effectively distinguishing the importance of different regions for recognizing expressions. However, this method is markedly dependent on neutral expressions in the data set. For the data, performance is good in the model training phase but poor in the validation and testing phases. An [11] proposed an adaptive model parameter initialization method based on the maxout network (MMN) linear activation function. This method effectively solves the overfitting problem and has poor algorithm robustness in the face of expression-independent factors. Liu [12] Through studied data samples and proposed deep features from significantly guided face regions by a conditional CNN-enhanced random forest algorithm (CoNERF) to suppress the effects of illumination, occlusion, and low resolution. However, this method is only for frontal facial expression image data extraction and retains considerable redundant information that could not be easily migrated. Occlusion occurs for faces in the facial expression database. To solve the occlusion problem in the recognition process, Ming Li [13] used an FER algorithm incorporating histogram of oriented gradient (HOG) features and improved KC-FDDL dictionary learning coefficient representation, combined with residual-weighted coefficient representation for expression classification. However, no training was performed on the deflected face sample data, resulting in poor recognition of the deflected face features. TAMFOUS et al. [14] used sparse coding and dictionary learning methods to study the time-varying shapes on the Kendall shape space of 2D and 3D landmarks, as well as addressed the effect of nonlinearity in the shape space on facial expressions. However, the accuracy of facial expression raw image feature recognition is not high. Jinghui Chu et al. [15] addressed the problem that expression features extracted by deep convolutional networks are susceptible to such factors as background and individual identity through lightweight CNN and attention models. However, the generalization ability of the model to recognize facial images is limited. To better reflect the goodness of the model through the loss function, Schroff [16] proposed the triplet loss function, which can better differentiate the details, although there is the problem of slow model convergence. Byoung [17–19] used the center loss function center loss to make the samples uniformly distributed around the center within the class and minimize the intra-class variation. However, computational efficiency is considerably low to recognize facial data features in real time. Lin [20] proposed focal loss function to improve the classification performance of the model by focusing the parameters to make the model focus considerably on difficult-to-classify samples. However, it cannot solve the problem of labeling face facial data samples.

The preceding analyses mainly focused on the facial expression detection algorithm and loss function, and extensive research has been conducted. However, limited studies have been conducted on data samples, which are deflected and judged separately with different angles, and weighted using probabilities to obtain the most likely recognition results. This study uses the CNN algorithm as basis to innovate the algorithm by using the concept of grayscale deflection weighting. The use of grayscale deflection weighting can synthesize the image feature information and perform expression recognition on dynamic and static images, improve the generalization ability of the model, reduce the probability of model prediction error, ensure data stability, and avoid overfitting caused by changes in the input data distribution.

The remainder of this study is organized as follows. Section 3 describes the system flow and image processing, and constructs a flowchart detailing the information related to the image to be tested in the system. Section 4 indicates that plots training is curved with confusion matrix through experiments and performs parameter tuning. This section also analyzes the results by experimentally using the same data set trained on different network models and compared. Lastly, Section 5 summarizes the study and provides the relevant conclusions.

3. Methodology

3.1 Image pre-processing

The image data set selected in this study is the CK+ data set. Each facial expression sequence in this data set represents one of the seven types of facial expressions (i.e., anger, surprise, disgust, enjoyment, fear, and sadness), as shown in Figure 1.

Fig. 1. CK+ data set

The image data set contains a large amount of image data. Given that there is a large amount of background data in the picture data interfering with the recognition results, this study will use face recognition region segmentation to segment the irrelevant regions. The specific flowchart is shown as follows.

Fig. 2. Flowchart of the image preprocessing

3.1.1 Face detection

For the face detection module, this study uses the Haar feature algorithm under the OpenCv framework to detect the face region. Haar feature extraction is mainly done by extracting features from the grayscale values of pixels. This method mainly performs face recognition by using the Cascade Classifier under OpenCv, which can detect face regions by inputting the Haar features to the Cascade Classifier [21].

3.1.2 Local binary pattern artificial feature extraction

Local binary pattern (LBP) is an operator used to characterize the local texture of images. LBP has significant advantages, such as rotational invariance and gray scale invariance. T. Ojala, M. Pietikäinen, and D. Harwood proposed LBP in 1994 for texture feature extraction, and the extracted features are local texture features of images [22].

The LBP algorithm is one of the first ones used in fingerprint recognition with good rotation invariance and grayscale invariance, and has been widely used in fingerprint recognition and image matching [23]. The two most common LBP methods are basic and ring LBP. The core point of the LBP algorithm is to compare the pixel value of the center point to be selected with the pixel value of the 8 neighbors of that point and set the marker symbol flag. If the point pixel value is higher than the point 8 neighbor pixel value, then the flag is set to 1. If the point pixel value is lower than the point 8 neighbor pixel value, then the flag is set to 0. Binary values of the 8 neighbor pixels (clockwise from the top left) are combined into a single 8-bit binary number, which is transformed thereafter into a decimal number to replace the middle pixel value. The LBP algorithm extends the 3 × 3 neighborhood of LBP to a neighborhood that can be represented by the parameters P and R, where P is the number of neighboring pixels and R is the radius of this neighborhood. Accordingly, the basic LBP algorithm that cannot considerably recognize large-scale texture features can be improved.

3.1.3 Size normalization

To avoid the subsequent expression recognition, numerous irrelevant information in the image leading to a certain error in recognition should be tested. Moreover, considering that the input side of CNN requires that the image is of fixed size, this study performs normalization operation on the subsequent image to be tested. The method used is to call the resize function in CV2. The specific method is a bilinear interpolation method, which is the method of four image points in the original image to represent the central image point, as shown in Figure 3.

Let the coordinates of the four vertices be as follows: $P1(x1, y1)$, $P2(x2, y2)$, $P3(x3, y3)$ and $P4(x4, y4)$. The coordinate of the synthesis point is $P(x, y)$. Thereafter, the equation of the point to be synthesized is as follows:

$$P(x,y) = P_1(x_1,y_1) * \omega_1 + P_2(x_2,y_2) * \omega_2 + P_3(x_3,y_3) * \omega_3 + P_4(x_4,y_4) * \omega_4 \quad (1)$$

where ω_i is the weight, and the weight of each point is related to the distance between the point to be solved and the diagonal point, in which ω_i is expressed as follows:

$$(x_2 - x)(y_2 - y) \quad (2)$$

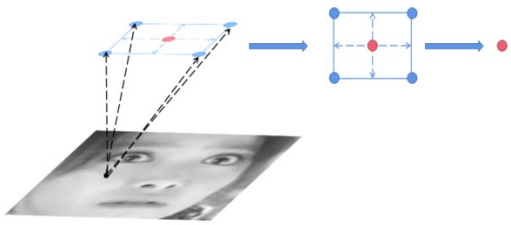

Fig. 3. Dimensional normalization flowchart

3.2 CNN

CNN is a typical feed-forward neural network [24]. It has a natural learning habit and expressive ability, given the fixed size of the input image. Features of the image can be extracted automatically. This neural network is often used to classify images.

The network often includes convolutional, pooling, and fully connected layers. A typical CNN structure is shown in Figure 4.

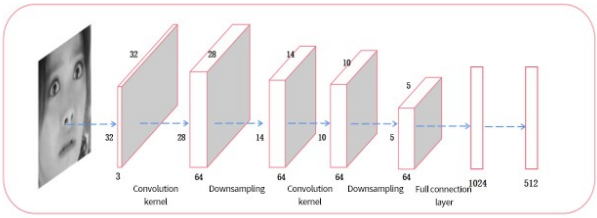

Fig. 4. Convolutional neural network results figure

3.2.1 Rollup

In general, a typical convolutional layer is connected to the input layer. Moreover, feature images from the input layer are fed to the convolutional layer, processed by convolutional kernels to obtain more abstract feature images, and passed to the next layer for processing. The equation for the convolution layer is as follows:

$$im_j^l = f(\sum_i im_j^{l-1} \Theta k_{ij}^l + b_j^l) \quad (3)$$

where im_j^l denotes the output of the j channel of the l layer of image im , f denotes the activation function $sigma$, Θ denotes the convolution operation, im_j^{l-1} denotes the output of the j channel of the $l-1$ layer of the image, k_{ij}^l denotes the convolution kernel, and b_j^l denotes the bias parameter.

3.2.2 Downsampling layer

As shown in Figure 4, the spatial resolution of the image is reduced after the operation of the sampling kernel is carried out. This operation reduces the dimensionality of the feature images transmitted in the previous layer and also ensures the invariance of the image size and features. This operation enables increasing the sample size of the image and ensures that the image is rotated, panned, leveled, distorted, scale changed, and rotated with reduced sensitivity of the output and image fidelity when the operation is performed. The equation for calculating the downsampling layer is as follows:

$$im_j^l = f(\beta_j^l down(im_j^{l-1}) + b_j^l) \tag{4}$$

where im_j^l is the output of the j channel of the l layer of the im image, f is the activation function $sigma$, and $down$ is the drop kernel operation, which moves the drop kernel to divide the feature figure im_j^{l-1} into multiple sub-features that do not overlap each other, and computes the multiple sub-features. The methods often used are maximization and averaging. For FER, the maximization method is typically able to obtain more features than the averaging method to recognize the subtle differences in facial expressions. Therefore, the maximization method is mainly used to process images. In particular, im_j^{l-1} denotes the output of the j channel of the $l-1$ layer of the image im , β_j^l denotes the downsampling kernel layer weight, and b_j^l denotes bias parameter.

3.2.3 Full connection layer

The main purpose of the fully connected layer is to completely connect the feature figures extracted by the neural network and output the two-dimensional feature images as one-dimensional features. This situation is shown in the following equation:

$$im_j^l = f(\omega^l im_j^{l-1} + b_j^l) \tag{5}$$

where im_j^{l-1} denotes the input layer image (i.e., image output by the previous downsampling layer), im_j^l denotes the output layer image, b_j^l denotes the bias function for the fully connected layer, and ω^l denotes the weight for the fully connected layer.

3.3 System process

The system flow of this study is shown in Figure 5. The main flow is as follows.

(1) The input of the image to be measured includes taking a static input and using OpenCv to call the live camera for each image frame.

(2) Pre-processing operations are performed on the image, such as graying the image to be measured and filtering noise data by median filtering.

(3) Haar-based face detection is performed on the image to detect the face part of the image to be measured.

(4) Haar-based face feature extraction is performed on the images.

(5) Image cropping is performed on the detected part of the face image to remove the image background interference.

(6) The image is subjected to a specification normalization operation, and the image size is normalized to a specification of 48 × 48.

(7) By using CNN for the combination of FER to be tested, different probability of distributions are obtained by deflecting the pictures by multiple angles. The final probability of distribution is obtained by weighting and adding these probability distributions; the one with the highest probability at this time is used as prediction result.

(8) Support vector machine (SVM) was used to classify facial features. Combined with CNN model, bagging election is carried out to classify difficult images in multiple combinations. Through bagging election, the integrated model is constructed to improve the prediction accuracy.

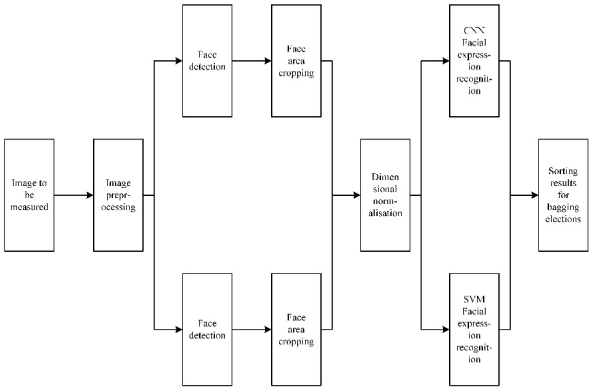

Fig. 5. System flowchart

4 Result Analysis and Discussion

4.1 Data Set Introduction

The CK+ data set was proposed by Lucey (Carnegie Mellon University) in the Cohn–Kanade data set, which has a total of 123 subjects. Dynamic videos in the data set contain emotion labels divided into consecutive video frames that indicate the expressions of the participants, with approximately 10,000 images of facial expressions from 123 models. These image sequences are continuous and have many similar images. The frame length range of this data set of facial expression sequences is 13 to 60 frames. The CK+ data set was collected in a laboratory setting. Each facial expression sequence in this data set represents one of the seven categories of facial expressions (i.e., anger, surprise, disgust, enjoyment, fear, and sadness) [25], as shown in Figure 6.

Fig. 6. CK+ data set

The distribution of category quantity of the CK+ data set is shown in Table 1.

Table. 1. Distribution of Category Quantity of the CK+ Data Sets

Types	Training Sets	Test Sets
anger	108	27
disgust	142	35
fear	60	15
enjoyment	166	41
sadness	67	17
surprise	199	50
normal	43	11

4.2 Experimental environment

Computer processors used in this experiment are as follows: AMD Ryzen 7 4800HS model, NVIDIA GeForce RTX2060, and 16 GBRAM is used as graphics card. The operating system is Windows10 (64 bit) and the software programming environment is Python 3.9.

4.3 Experimental results and analysis

On the premise of running CNN algorithm for expression recognition on the CK+ data set, this study proposes an innovative algorithm to obtain different probability distributions by using images deflected by multiple angles and weighing these probability distributions to obtain the final probability distribution. At this time, the one with the highest probability is used as predictive result. Examples of static and dynamic expression results recognized by the proposed algorithm are as follows.

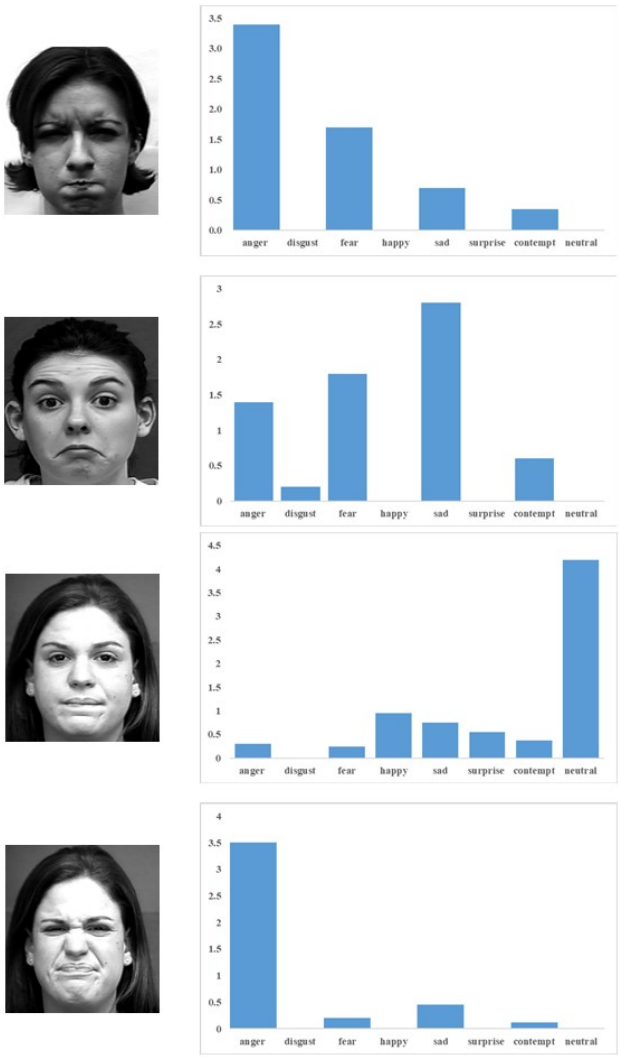

Fig 7. Static example result diagram

Fig 8. Dynamic example result figure

Note that based on the observation of the preceding figure, the improved algorithm can better recognize the expression of the image to be tested. Thereafter, the entire data set was predicted by this study. As the training continues, the training curve fluctuates constantly. However, the accuracy of the training and verification sets continuously improves gradually. Lastly, both curves reach a stable state, and the training curve obtained is shown in Figure 9.

After the training is completed, the final test is validated in the test set. This study uses the confusion matrix to evaluate the effect of the model on the test set. The results of its confusion matrix are shown in Figure 10. The recognition rate of happy is 92%, while the lowest recognition rate of scared is only 56%. This result is possibly caused by the uneven number of different expressions. During the identification process, sad, scared, angry, and disgust are considerably difficult to identify.

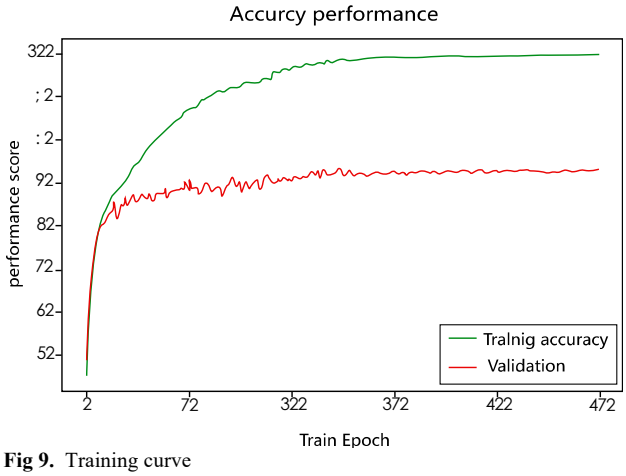

Fig 9. Training curve

The confusion matrix shows that the accuracies in the test set of the seven expressions of angry, disgust, fear, happiness, sadness, surprise, and neutral are 67%, 67%, 56%, 92%, 61%, 82%, and 73%, respectively.

After subsequent parameter tuning and code optimization, the model achieved an accuracy of 97.85% on the CK+ data set.

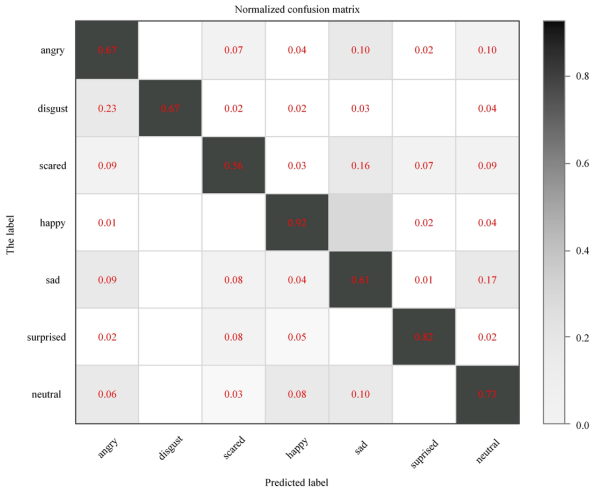

Fig 10. Confusion matrix

Table 2. Comparison of the recognition rates

Literature	Algorithms	Accuracy (%)
26	Fusion Network for Classification	97.35
27	Cross-connect LeNet-5 network	94.37
28	Deep Neural Network	96.8
29	Learning DeepSparse Autoencoders	95.79
This study	Improving CNN networks	97.85

To verify the accuracy of the improved CNN method in this study on the CK+ data set, several algorithms from the literature are used for comparison with the current research. Table 2 shows that the improved CNN method has a significant improvement in recognition rate compared with other expression recognition methods. The literature [26] has extracted traditional manual features, but the loss of features affects the classification with an accuracy of 97.35%.

To verify the accuracy of the improved CNN method in this study on the CK+ dataset, several algorithms from the literature are used for the comparison with this study. As can be seen from Table 2, the improved CNN method has a significant improvement in recognition rate compared with other expression recognition methods. The literature [26] extracted traditional manual features, but the loss of features affects the classification with an accuracy of 97.35%.

The literature [27] has relatively few useful features that have an impact on classification in low-level feature extraction. Therefore, the accuracy of the results is relatively low. On the basis of the LeNet-5 network, a new CNN structure is designed by introducing the cross-connectivity method and applying it to FER. Moreover, the LeNet-5 structure is applied to perform poorly in expression recognition in a handwritten digital library. The literature [28] used a combination of traditional methods and deep learning, using units of the traditional feature descriptor LBP for supplementary training. However, the current research uses a CNN improvement algorithm for feature extraction with high accuracy. The literature [29] utilized a combination of high- and low-level methods for expression recognition, with complex network structure design units, reduced completeness of the feature extraction process, and low accuracy rates. The preceding comparative analysis indicates that the simplification of the number of convolutional layer modules in the network model and the removal of the final connection layer in this study substantially improve the recognition speed and authenticity of expression recognition.

5. Conclusions

The concept of gray deflection weighting on the basis of the CNN algorithm was innovated to explore deep learning in the application of facial expression and HCI and other fields to solve the problem of low recognition rate of facial expression. By using OpenCv combined with gray distribution probability for expression recognition, linear weighting was used to obtain the final FER system. The accuracy of FER was improved. After the experiment and summary of the system, the following conclusions are drawn.

(1) The use of the gray-scale deflection weighting concept can synthesize image feature information, improve the generalization ability of the proposed model, reduce the model prediction error probability, ensure data stability, and avoid the overfitting problem caused by the change of input data distribution.

(2) The proposed model combines algorithmic innovation by invoking OpenCv to perform expression recognition on dynamic and static images, and eliminating the risk of fraud that may occur in previous literature by performing expression recognition on static images only.

(3) An accuracy of 96.05% was achieved on the restricted CK+ data set using an improved CNN algorithm. The results are better than those of many existing mainstream FER methods.

The FER algorithm based on image deflection angle weighting is proposed in this study, specifically by combining simulation and theory. The proposed algorithm is markedly accurate and has a certain reference for the subsequent development of facial expression recognition. Angry and disgust expressions are prone to miscategorization because they have similar eyebrow features, and disgust with a wrinkled mouth. People who wear glasses are often wrongly classified as angry or scared. Given that dark frames are often confused with frowns that characterize both expressions, happy and surprise are also a cause of confusion because eyebrows are raised at similar angles. These pieces of evidence suggest that human expression is substantially complex. Moreover, expression recognition is a complex and ambiguous study. Hence, the proposed model needs additional refinement.

Acknowledgements

This work was supported by the Natural Science Foundation of Zhejiang Province, China (LTY20E050002).

References

1. Chau M. H., Jacobs G. M., "Applied Linguistics, language guidelines, and inclusive practices: The case for the use of who with nonhuman animals". *International Journalof Applied Linguistics*, 31(2), 2021, pp.301-303.
2. Li S., Deng W., "Deep facial expression recognition: A survey". *Transactions on Affective Computing*, 2020, pp.99.
3. Naidu P. R., Sagar S. P., Praveen K., Kiran K., Khalandar K., "Stress Recognition Using Facial Landmarks and Cnn (Alexnet)". In: *International Conference on Applied Mathematics, Modeling and Simulation in Engineering*, Hyderabad, India: IOP Science, 2021, pp.10.
4. Li W. T., Luo X. S., Meng Z. M., Chen J., "A study on face expression recognition combining improved RepVGG-A0 network and relabeling". *Modern Electronics Technology*, 45(20), 2022, pp.69-74.
5. Zhou J., Ma M. D., "Face Expression Recognition Based on Improved ResNet Network". *Computer Technology and Development*, 32(01), 2022, pp.25-29.
6. Chen J., Shan S. G., He C., Zhao G. Y., Pietikainen M., Chen X., Gao W., "WLD: A robust local image descriptor." *Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 2009, pp.1705-1720.
7. Tan X. H., Li Z. W., Fan Y. C., "A multi-scale detail enhancement-based facial expression recognition method". *Journal of Electronics and Information*, 41(11), 2019, pp.2752-2759.
8. Alpaslan N., Hanbay K., "Multi-resolution intrinsic texture geometry-based local binary pattern for texture classification". *Access*, 8, 2020, pp.54415-54430.
9. Hanbay K., Alpaslan N., Talu M. F., Hanbay D., Karci A., Kocamaz F. A., "Continuous rotation invariant features for gradient-based texture classification". *Computer Vision and Image Understanding*, 132, 2015, pp.87-101.
10. Zhou H. P., Zhang D. Y., Sun K. L., Gui H. X., "A face expression feature extraction method based on ENM-Gabor difference weights". *Computer Applications and Software*, 37(03), 2020, pp.184-189+212.
11. Fengping An and Zhiwen Liu. "Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM". *The Visual Computer*, 36(3), 2020, pp.483-498.
12. Liu Y. Y., Yuan X. H., Gong X., Xie Z., Fang F., Luo Z. G., "Conditional convolution neural network enhanced random forest for facial expression recognition". *Pattern Recognition*, 84, 2018, pp.251-261.
13. Li M., Peng X. J., Wang Y., "Face expression recognition based on improved dictionary learning with sparse representation". *Journal of Systems Simulation*, 30(01), 2018, pp.28-35+44.
14. Tamfous A. B., Drira H., Amor B. B., "Sparse coding of shape trajectories for facial expression and action recognition". *Transactions on Pattern Analysis and Machine Intelligence*, 42(10), 2020, pp.2549-2607.
15. Chu J. H., Tang W. H., Zhang S., Lv W., "An attention model-based algorithm for facial expression recognition". *Advances in Lasers and Optoelectronics*, 57(12), 2020, pp.205-212.
16. Schroff F., Kalenichenko D., Philbin J., "Facenet: A unified embedding for face recognition and clustering". In: *Conference on Computer Vision and Pattern Recognition*, Boston, USA: IEEE, 2015, pp.815-823.
17. Ko B. C., "A brief review of facial emotion recognition based on visual information". *Sensors*, 18(2), 2018, pp.401.
18. Kim S., Nam J., Ko B. C., "Facial Expression Recognition Based on Squeeze Vision Transformer". *Sensors*, 22(10), 2022, pp.3729.
19. Kim S., Jang I. S., Ko B. C., "Image Registration Between Real Image and Virtual Image Based on Self-supervised Keypoint Learning". In: *Asian Conference on Pattern Recognition*, Jeju Island, Korea: Springer, 2022, pp.402-410.
20. Lin T. Y., Goyal P., Girshick R., He K. M., Dollar P., "Focal loss for dense object detection". In: *International Conference on Computer Vision*, Venice, Italy: IEEE, 2017, pp.2980-2988.
21. Hu Y. F., Hu Y. B., Li Q., Geng D. D., "Research on face detection and tracking recognition system based on video surveillance". *Computer Engineering and Applications*,52(21), 2016, pp.1-7+35.
22. Song Y. P., Huang H., Ku F. L., Fan D. D., "Face recognition algorithm based on MB-LBP and tensor HOSVD". *Computer Engineering and Design*, 42(04), 2021, pp.1122-1127.
23. Yao L. P., Pan Z. L., "Research on face recognition method based on improved HOG and LBP algorithms". *Optoelectronics Technology*, 40(02), 2020, pp.114-118+124.
24. Zhang T., Yang J., Song W. A., Guo Y. Y., "Improved convolutional neural network model design. Improved convolutional neural network model design method". *Computer Engineering and Design*, 40(07), 2019, pp.1885-1890.
25. Lucey P., Cohn J. F., Kanade T., Saragih J., Ambadar Z., Matthews L., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression". In: *Computer Society Conference on Computer Vision and Pattern Recognition-workshops*, San Francisco, USA: IEEE, 2010, pp.94-101.
26. Zeng G., Zhou J., Jia X., Xie W. C., Shen L. L., "Hand-crafted feature guided deep learning for facial expression recognition". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, Xi'an, China: IEEE, 2018, pp.423-430.
27. Lin X. Z., Jiang M. Y., "Facial expression recognition with cross-connect LeNet-5 network". *Acta Automatica Sinice*, 44(1), 2018, pp.176-182.
28. Li J., Lam E. Y., "Facial expression recognition using deep neural networks". In: *International Conference on Imaging Systems and Techniques*, Macau, China: IEEE,2015, pp.1-6.
29. Zeng N., Zhang H., Song B., Liu W. B., Li Y. R., Dobaie A., "Facial expression recognition via learning deep sparse autoencoders". *Neurocomputing*, 273, 2018, pp.643-649.